REAL-TIME SOURCE ACTIVATED SHAPING

BACKGROUND

1. Field

This disclosure relates to real-time transmission over data networks, more particularly to methods and devices used to improve quality of service for real-time transmission over data networks.

2. Background

As more sophisticated equipment has become available, transmission of data from real-time sources such as voice and video has become more prevalent. Data networks are often designed to accommodate the maximum bandwidth required. However, the network traffic may 'burst' up to that capacity, but more often runs at a lower rate. This provides excess capacity in data networks to transmit such things as video and voice. Enterprises, such as companies and government entities, have found the use of data networks for voice transmission, especially, to be a cost-effective alternative to using the public switched transmission network (PSTN).

However, the use of data networks for real-time transmissions such as voice and voice raises concerns about the quality of service (QoS). Network access providers typically offer a range of service levels to enterprises for corresponding fees. The network may be 'purchased' by the enterprise as a virtual circuit, which is a logical circuit created to ensure reliable communication between network devices.

One concern about QoS is congestion on the network. Even though the network may have excess capacity at some times, at other time it may be 'full' and there may be delays in transmission. If the transmissions are data packets or modules, there is generally not a problem except the sender and receiver's inconvenience. However, if the transmissions are real-time data packets or modules, the delay may be fatal to the overall transmission, since the packets or modules have to arrive with enough time to be decoded into a sequential data stream. Delayed packets or modules will cause the sequential order to be disrupted and will adversely affect the reception of the real-time transmission.

A possible solution is to have a VC dedicated to voice. However, this would typically require an enterprise or user to have two VCs, one dedicated to voice and one for data. Even if the enterprise or user only used one VC, the costs would be greater. It would be preferable for the voice and data traffic to share a VC but without sacrificing quality of service for the voice, and still have a reasonable cost.

SUMMARY

One aspect of the disclosure is a network device. The network device includes an input port to receive input data and a transmission port to transmit data. The device also includes a detector to detect real-time input data and a controller to set the maximum transmission rate equal to a first traffic rate when the detector detects real-time input data. A method of managing traffic in the presence of real-time data is also included. The method detects the real-time data and then sets a maximum transmission rate to a first traffic rate.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention may be best understood by reading the disclosure with reference to the drawings, wherein:

Figure 1 shows an embodiment of a network having a virtual circuit, in accordance with the invention.

Figure 2 shows an embodiment of a network device, in accordance with the invention.

Figure 3 shows an embodiment of a method of managing traffic in the presence of real-time data, in accordance with the invention.

Figure 4 shows an embodiment of a method for detecting real-time data, in accordance with the invention.

Figure 5 shows an alternative embodiment of a method for detecting real-time data, in accordance with the invention.

DETAILED DESCRIPTION OF THE EMBODIMENTS

The issues raised by the use of data networks for real-time transmissions have given rise to some possible solutions other than the dedication of a virtual circuit (VC) to voice, discussed above. Many network providers offer two levels of bandwidth or

'information rate.' One level may be referred to as a 'committed information rate (CIR),' being the maximum bandwidth guaranteed for a particular VC. This may also be referred to here as the first traffic rate. A second level of bandwidth may be made available to the users on a non-guaranteed basis, sometimes referred to as an 'excess information rate (EIR).' This may also be referred to here as the second traffic rate. This may be viewed as 'best effort' data, where the network will make best efforts to get the data through.

The use of EIR allows users to occasionally exceed their CIR, allowing them a higher maximum transmission rate, defined here as the rate at which their information actually flows across the network, between these two rates. Networks may mark EIR traffic to notify the network that the data may be discarded if necessary. For example, in Frame Relay networks, the data would be marked DE, for discard eligible. Usually, the network will only discard data to alleviate congestion. In some cases, for example, EIR data may be discarded to preserve the CIR guarantees made for the VCs.

One possibility for users wanted a guaranteed data rate for their real-time data is to pay for a committed information rate that is sufficient for both their voice and data traffic, with no excess information rate. This would guarantee that all data would get through and the quality of service would be maintained. A problem with this approach lies in both the cost and the inability of customers to have bursts of traffic over their CIR.

Another possibility is to use a CIR sufficient for the real-time traffic and some percentage of the data. The EIR would be available for best efforts data as required. All the data traffic would be marked as discardable and the user would reduce the total traffic rate offered towards the CIR in response to congestion conditions in the network, including congestion notifications. A possible risk here is that the user is offering data for discard ahead of other users, and any delays in congestion notifications may allow congestion to occur, resulting in discards being made that might not have been necessary. Also, the network may not take discard markings made by the user into account, and therefore may not protect the real-time traffic as desired.

Application of this invention may result in a better solution to the problem. Application of the invention will generally allow users to detect or otherwise identify

real-time data in the VC and allows the user to reduce the maximum transmission rate to the CIR while real-time data is present. This ensures that all traffic is within the CIR when real-time data is present, and should have guaranteed delivery. It also allows full use of the EIR when no real-time traffic is present. This will also mitigate the users' costs, as they can pay for the necessary level of CIR as they determine their needs, not based upon a guaranteed level of delivery higher than they need when real-time traffic is not present.

Turning now to Figure 1, an overall network configuration can be used as context for understanding the invention. This is only intended as an example, and is not intended to limit application of the invention in any way. The network 10 may comprise several individual machines, or network devices, such as 12 and 14 and may include smaller networks such as 16. A virtual circuit may be set up between machines 12 and 14, which will define a logical circuit between these two machines across the network 10, which includes networks 16 and 18.

The machines 12 and 14, as well as any intervening machines may be any type of equipment that communicates across the network, including computers, workstations, routers, multi-channel adapters (MCAs), multi-channel concentrators (MCCs), etc. These network-capable machines will be referred to here as network devices.

One embodiment of a network device in accordance with the invention is shown in Figure 2. The network device 20 has an input port 22 and a transmission port 24a. It may also have transmission port 24b. Similarly, it may transmit and receive through the same port. A detector module 26 is capable of detecting the presence of real-time data, which will be discussed in more detail later. A controller 28 is capable of reducing the maximum transmission rate for the device to the first traffic rate in response to the presence of real-time data. The controller and the detector may reside on the same component 30, such as a microcontroller or a processor. It may also be an application specific integrated circuit (ASIC), a general-purpose processor or central processing unit, or a digital signal processor, as examples.

One embodiment of a method for managing traffic in a network device is shown in Figure 3. At 32, real-time data is detected. As will be discussed in more detail, the real-time data may actually be data or may be real-time traffic. Traffic

would include packetized data that may not be individually analyzed to determine if the data inside the packet is actually real-time data. When real-time data is detected, the maximum transmission rate is reduced to a first traffic rate, referred to as the CIR in the above discussion.

Optionally, the method may include a recovery process in which the cessation of real-time data is also monitored. Upon cessation of real-time data, or the absence of real-time data, the network device would then allow the maximum transmission rate to exceed the first rate. Generally, the maximum transmission rate would be bounded by the first transmission rate and a second transmission rate, the two rates being the CIR and the EIR, respectively. However, the recovery process is not necessary for application of the invention.

The detection of real-time data at 32 may be accomplished in many different ways. For example, the packets of data flowing through, from or to a network device could be examined for particular characteristics that identify them as real-time data. For example, the data inside the payload may have particular characteristics or profiles that indicate that it is real time data. Alternatively, the header may identify the source of the data, which in turn may be known to the network device as a real-time source, such as a voice coder/decoder (codec) or a video terminal attached to the network. This embodiment is not limited to any particular place in the VC. It may be applied at the originating end, the receiving end or by any device in between.

Similarly, if the real-time data were identified on a packet basis, the optional recovery process may include a timer in the network device. Each time a real-time packet is identified, the timer is reset. The controller would then monitor the timer for expiration. If no real-time packets are identified within the period of time set by the timer, the controller releases the maximum transmission rate and allows it to exceed the first traffic rate.

Other options for detection of real-time data may be more specific to the location of the device in the network. For example, an originating or receiving end of the VC may have specialized equipment such as voice or video codecs that identify the data as real-time. This equipment is referred to here as sources of real-time content. In the case of a network device electrically coupled to a real-time source, the real-time source may indicate its impending transmission of data. For example, the video or voice codec may take control of a data bus, indicating that it is about to

transmit voice or video data through the codec's port on the bus. This would allow the network device to also become aware that real-time data transmission is imminent and to set the maximum transmission rate to the first transmission rate. This is shown in more detail in Figure 4.

Process 32 from Figure 3 has been expanded to include monitoring a port electrically coupled to a source of real-time data at 40. A signal is received at 42 that indicates the real-time data transmission is about to occur. In response to the signal, the maximum transmission rate is set to the first transmission rate at 44.

In this particular embodiment, the determination of whether real-time data is present is altered slightly from Figure 3. Here, the determination is whether a second signal indicating cessation of the real-time data transmission is received at 36. This then determines the point at which the maximum transmission rate is allowed to exceed the first information rate.

The specifics of implementation are left largely to the system designer based upon the features and capabilities of the networks. For example, a network may use some source of resource reservation, such as the resource reservation protocol (RSVP) used in Internet Protocol (IP) networks, which may be carried on Frame Relay networks. In this case, shown in Figure 5, the detection of real-time data will be the reception of the resource reservation request at 50 indicating the that maximum transmission rate should be reduced to the first traffic rate at 52. The first traffic rate may be 'announced' by the reservation request, or it may be predetermined for the network devices receiving the request.

Recovery of the maximum transmission rate in this type of scheme generally involves receiving some sort of release message that indicates that the resources are no longer needed. This would then trigger the release of the maximum transmission rate and allowing it to exceed the first traffic rate.

There is no specific limitation as to the type of network to which this invention can be applied. Currently, VCs are more common on Frame Relay, Asynchronous Transfer Mode, Internet Protocol tunnels, and local area network links, as examples. However, each type of network generally has an analogous structure or capability to those defined above and this invention could be applied to any network that includes transmission of real-time data.

The network devices may include the capabilities set forth above, or they may be upgraded to include those capabilities. Generally, the upgrade will be to the software code that operates the various types of devices. In that instance, an article will include machine-readable code that, when executed, causes the machine to perform the methods of the invention.

Thus, although there has been described to this point a particular embodiment for a method and apparatus for real-time source activated shaping, it is not intended that such specific references be considered as limitations upon the scope of this invention except in-so-far as set forth in the following claims.